REMARKS

Amendments to the Specification

The paragraph at page 8, line 19 through page 9, line 2 has been amended to remove the hyperlink references that were objected to by the Examiner.

Amendments to the Claims

Claims 6 and 15 have been cancelled. Claims 19 and 20 have been added and are essentially the text of claims 6 and 15 rewritten in independent form but amended to contemplate a subject protein sequence or subsequence instead of a subject genome sequence. No new matter has been added.

Claims 1, 10 and 13 have been amended to specify that the formed vector provides a uniform representation of the subject genome sequence. Support for these amendments can be found in the specification at least on page 4, lines 17-19 and at page 7, lines 7-10 as originally filed.

Claims 3 and 12 have been amended to clarify that the step of repeating the comparison and forming steps for each subject genome sequence comprises using the same set of known biological fragments as was used for the first comparison. Support for this amendment can be found in the specification at least on page 9, lines 6-9 as originally filed. Claims 3 and 12 have also been amended to state that the provided vectors are of uniform length. Support for these amendments can be found in the specification at least on page 4, lines 17-19 and at page 7, lines 7-10 as originally filed.

No new matter has been added.

Applicant's Invention

Biological sequence research has reached a level where high-throughput processes can generate thousands of hypothetical genes that have not been assigned a putative function. It is well known that the function of a protein is dictated by its amino acid sequences which determine the protein's structure, and thus its interaction with the environment. However, biologists have identified four levels of structure which can influence the protein's function. The primary structure is the sequence of amino acids. The secondary structure is the presence or absence of

small "sub-folds". The tertiary structure is the final three-dimensional shape. The quaternary structure includes the complexes formed with other proteins. Given one level of structure, it is not an easy task to predict the next levels. Thus, it is desirable to have a technique other than a complete sequencing to determine the three-dimensional structure and ultimately the protein function. Otherwise, sequencing would be required for all of the thousands of genes produced by the aforementioned high-throughput methods.

The present invention utilizes known biological fragments to provide information about the unknown, or subject genome sequence. The known biological fragments may be taken from known and publicly available databases which contain a protein's primary structure annotated with their secondary and tertiary structures. One use of the present invention is to use this known data to build models for known protein structures, and then to automatically annotate new proteins according to the models. See Spec. page 4, lines 1-16.

A method of the present invention comprises the steps necessary to form a uniform vector representation of a subject genome sequence based on a comparison of the subject genome sequence to a predetermined number of known biological fragments. The first step is to provide a set of predetermined number of known biological fragments, for example a motif database. Then, a comparison is made between the subject genome sequence and each known biological fragment in the provided set. For each known biological fragment in the set, the number of times that the fragment occurs in the subject genome sequence becomes the respective vector element for that known biological fragment. Finally, the set of vector elements (defined by the foregoing comparisons) forms a vector having a length equal to the number of known biological fragments provided in the set and used in the comparisons. This vector is referred to as a uniform representation of the subject genome sequence because the comparison of any other subject genome sequence to the same set of known biological fragments will result in a respective vector of the same length. That is, all of the resulting vectors will have the same length. See Spec. page 4, lines 17-26, page 7, lines 3-10, page 8, lines 4-8, and page 9, lines 3-19.

Where the resulting vectors have the same length, analyses of interest (e.g. clustering, classification, and indexing, etc.) can be applied to (across all) the subject genome sequences in an effective and reliable manner. See Spec. page 7, lines 13-16 and lines 24-29 and page 10, lines 5-16.

Another aspect of Applicant's invention is an apparatus for implementing the above-described method, for example, a computer system.

With Respect to the Drawings

Accompanying this Amendment are formal drawings for filing in the subject application.

Sequence Non-Compliance

The Office Action at hand states that the application fails to comply with the requirements of 37 C.F.R. § 1.821 through 1.825 because FIG. 3 contains an amino acid sequence. The Office Action further states that Applicants must provide a Sequence Listing, a computer readable format of the Sequence listing, and a statement that the two are identical, and use a sequence identifier, either in the drawing or in the Brief Description of the Drawings.

The protein sequence shown in FIG. 3 is not an actual amino acid sequence contemplated by Applicants. Rather, it is a fictitious sequence of amino acids which represents a possible subject input character string that one might be researching and it is used simply for the purpose of illustrating the present invention. The letters used in this fictitious sequence are not meant to represent any specific amino acids, but rather they are variables which could represent any amino acid of the user's choosing. For example, the fictitious sequence begins "SVYDAAAQL." Referring to Table 3: List of Amino Acids in MPEP § 2422, there is no amino acid represented by SVY, DAA, or AQL.

The purpose of submitting a Sequence Listing is to present nucleotide and amino acid sequence data in a manner that is consistent across all patent applications to make searching and examination easier. See MPEP § 2420. However, Applicants do not contemplate any specific sequence for this application and have instead used a representative hypothetical sequence as shown in FIG. 3. Thus, it is respectfully requested that the requirement that the sequence represented in FIG. 3 comply with the sequence listing rules of 37 C.F.R. § 1.821 through 1.825 be withdrawn as not furthering the purpose of Sequence Listings. Alternatively, Applicants propose to simply delete the character string "SVYDAAAQLTADVKKDLRDSWGVAL..." from FIG. 3 and replace in its stead a rectangle symbolic of a protein sequence. Such an amendment will not introduce any new matter. If upon further examination it is required of Applicant to

submit a Sequence Listing and a computer readable format for the fictitious sequence shown in FIG. 3, then Applicants will so provide.

Objection to Specification

The Specification has been objected to for containing embedded hyperlinks at page 8, lines 22-23 and lines 28-29. The hyperlinks have been deleted from the specification by this Amendment. Thus, it is respectfully requested that the objection to the specification be removed.

Rejection of Claims 1-9 under 35 U.S.C. § 101

Claims 1-9 have been rejected under 35 U.S.C. § 101 as being directed to non-statutory subject matter. In support of the rejection the Office Action states that the fixed length vector representation of a subject genome sequence as claimed is a "representational event of manipulated data that is determined by a mental process that solves a mathematical algorithm."

MPEP § 2106 (IV)(B)(2)(b)(ii) states that a process which merely manipulates an abstract idea or performs a purely mathematical algorithm is non-statutory subject matter. Claim 1 recites more than a mathematical algorithm or representational event. Claim 1 comprises the step of providing a set of known biological fragments. There are no mathematical calculations as part of this step. The specification teaches at page 8, line 19 through page 9, line 2 that the set of known biological fragments can be published databases or motifs of protein sequences, or can be a newly created motif database from any protein database which has been labeled according to some parameter (e.g. structure). In fact, it is desirable to create motif databases that are specific to the subject genome sequence because such a database will result in the generation of more meaningful vector representations than if a general or random database is used because such a database may not contain many motifs which occur in the subject genome sequence. Thus, with this first step of Claim 1, the claimed process does not consist "solely of mathematical operations, i.e. converting a set of number into another set of numbers..."

To further clarify the subject matter to which the claims are directed, base Claim 1 has been amended to recite "a method for analyzing genome sequences." As now claimed, the method includes the steps of "...providing a set of known biological fragments, ... comparing the respective representations of each known biological fragment ... to a subject genome sequence...;

... forming a vector ... such that the formed vector provides a uniform representation of the subject genome sequence; and providing the formed vector for use as input to a desired analysis, the uniform representation provided by the formed vector enabling the formed vector to serve as normalized input." Thus, the claimed method is believed to be directed to statutory subject matter within the meaning of 35 U.S.C. § 101 (i.e. "a new and useful process") for which a patent may be obtained.

Support for these claim amendments is found at least on Spec. page 6, line 24 - page 7, line 29 as originally filed. No new matter is introduced.

Claims 2-9 are dependent on Claim 1 and follow the foregoing. As such, it is believed that the rejection of Claims 1-9 is overcome.

Claims 1-18 have been rejected under 35 U.S.C. § 101 as lacking patentable utility because "the application lacks specific and substantial utility of the fixed length vector representation of a subject genome sequence as claimed." The Office Action argues that no relationship can be derived from the data collected for forming the vector concerning possible genome information. It is further stated that the claimed fixed length representation is not supported by a specific asserted utility and that the suggested analyses of classifying, indexing or clustering have no stated purpose.

Base Claims 1 and 10 have been amended to recite a method and apparatus, respectively, for analyzing genome sequences. One element of the claimed method and apparatus is a "formed vector [that] provides a uniform representation of the subject genome sequence." Also, as now claimed, the uniform representation of the formed vector provides or serves as normalized input to a desired analysis (especially automated analyses) of the subject genome sequence. See Specification page 7, lines 3-16 and lines 24-29 as originally filed. No new matter is being introduced.

Thus, the invention as now claimed provides mathematical (vector based) manipulation of data determined from biological fragment or sequence presence. Such is said to "..[lead] to specific and substantial utility." See page 4, middle paragraph, of the Office Action at hand.

Dependent Claims 2-9 and 11-18 include the elements and features of the respective base Claims 1 and 10. The foregoing arguments thus equally apply to Claims 2-9 and 11-18.

Furthermore, as stated above in the description of Applicants' invention, the claimed set of known biological fragments are selected because certain genomic characteristics are already known about them. For example, the secondary and tertiary structures of the known biological fragments can be the genomic characteristic that is already known. Thus, once a vector for the subject genome sequence is obtained which reveals the presence and quantity of known biological fragments, information regarding the secondary and tertiary structure of the subject genome sequence may be inferred. See the specification at page 4, lines 10-16 as originally filed. As discussed above, these structures will determine functional characteristics of the subject biological sequence, which is one of the primary goals of biological sequence research. Thus, the specific and substantial utility of the present invention is the determination of characteristics of a subject genome sequence by comparing it to a database of known biological fragments for which the characteristics are already known. By representing the subject genome sequence as a vector which quantifies the frequency of which known biological fragments occur, these characteristics can then be readily obtained and utilized in further analysis of interest.

Given the specific and substantial utility of being able to identify characteristics of an unknown subject genome sequence by representing it as a vector which quantifies the presence of known biological fragments, for which the characteristics are known, and thus obviating the need for further laboratory analysis of the subject genome sequence, it is respectfully submitted that the invention of the present application satisfies the 35 U.S.C. § 101 utility requirement. It is thus believed that the § 101 rejection of Claims 1-18 is overcome. Acceptance is respectfully requested.

Claims 1-18 have been rejected under 35 U.S.C. § 112, first paragraph, as not being "supported by a specific, substantial and credible utility such that one skilled in the art would not know how to use the claimed invention or alternatively a well established utility."

As argued above, now amended Claims 1-18 are believed to recite an invention that is supported by a specific, substantial and credible utility as well as a well established utility (namely, normalization of data for input to known analyses of interest in biotechnology). Thus, Claims 1-18 as now amended are believed to overcome the rejection under § 112, first paragraph. Acceptance is respectfully requested.

Rejection of Claims 1-18 under 35 U.S.C. § 112, first paragraph

Claims 1-18 have been rejected under 35 U.S.C. § 112, first paragraph, as "containing subject matter which was not described in the specification in such a way as to enable one skilled in the art to which it pertains, or with which it is most nearly connected, to make and/or use the invention." In support of this rejection, the Office Action argues that "the instant application fails to provide guidance to one of ordinary skill in the art for generating a fixed length vector from a predetermined set of biological fragments or sequences as recited in claims 1 and 10." The Office Action further argues that, although the specification states that the set of known biological fragments can be selected from published sequences, motif or database, the specification does not disclose how to generate a set of known biological sequences that would lead to the construction of a meaningful vector representation.

Applicants' invention is one of "predictive analysis" (specification page 4 , line 3). Thus, the methods described in this application are for using what is known to predict what is unknown. Thus, it is readily evident to one ordinarily skilled in the art that the selection of known biological fragments must be made according to what one wishes to predict about the unknown subject genome sequence. An example is given in the specification on page 4, lines 10-16. In that example, it was desired to learn about the unknown subject genome sequence's structure. Thus, it is stated that the set of known biological fragments to be used here is a database which contains protein domain sequences (primary structure) annotated with their secondary and tertiary structure.

If, instead of protein structure, one wished to determine a different characteristic of the subject genome sequence, then one would simply select a set of known biological fragments for which that characteristic is also known. Comparison of that set of known biological fragments with the subject genome sequence according to the present invention would thus reveal the desired characteristics.

The specification goes on to detail how to create or obtain a comparison database (a predetermined set of known biological fragments/sequences) at page 8, line 9 - page 9, lines 2 as originally filed. In particular, the passage at page 8, lines 9-11, specifies that the desired set of known biological fragments/sequences is formed of "short, highly conserved regions in related protein domains." Specific published databases (e.g. BLOCKS, Emotif and PRINTs) for use in

the invention are identified at specification page 8, lines 19-24. An example using the BLOCKS database is described beginning at specification page 11, line 13. Specification page 8, lines 23-23 and page 11, lines 18-19 suggest the use of the whole published database "... as the working predefined set/comparison database 17" from which to create the feature (claimed) vectors.

Specification page 8, line 25 - page 9, line 2 as originally filed discloses an alternative to the foregoing for obtaining a comparison database (predetermined set). The alternative set forth suggests the creation of "a new motif database from any protein database which has been labeled according to some parameter (e.g. structure)." This part of the Specification goes on to describe the use of multiple alignment software "to find short multiply aligned ungapped sequences...," and then suggest the collecting of statistics about these sequences in a matrix to form the motif (comparison) database. Specification page 8, lines 17-18 describes how to represent a working motif in a matrix.

The statement on Specification page 8, line 29 - page 9, line 2 suggests optimization or improvement may be had by creating a comparison database that is specific to the proteins of interest. This passage recites that "more meaningful feature vectors 23 may be obtained" relative to the use of a more general database. This is not to be interpreted to mean that no meaningful vector will result from the use of a more general database in contrast to the arguments on page 6 of the Office Action at hand.

Thus, one skilled in the art, using the examples, specified published databases, and other descriptions recited in the Specification as originally filed (and summarized above), would be sufficiently enabled to practice the claimed invention, i.e. provide the claimed set of known biological fragments and form the claimed vectors. As such, it is respectfully requested that the Examiner withdraw this rejection under 35 U.S.C. § 112, first paragraph.

Claims 1-18 have been rejected under 35 U.S.C. § 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

Claim 1 and those claims dependent on it have been rejected due to lack of clarity of the term "biological fragment". In the specification at page 3, lines 2-3 and page 8, lines 10-11, the terms "region," "blocks," "motifs," and "probabilistic templates" are used for biological fragments. On Specification page 8, lines 12-14, it is enumerated that the biological fragments

are amino acids if the subject genome sequence is a protein and it is a nucleotide if the subject genome sequence is DNA. Thus, these discussions of generating the set of biological fragments clearly defines the metes and bounds of the claimed "biological fragments."

Claims 3 and 12, and claims dependent thereon, are rejected as vague and indefinite because it is unclear if the same set of biological fragments as in Claim 1is utilized for the vector construction. These claims have been amended to indicate that the same set of biological fragments is indeed used for the generation of additional vectors. Claims 4, 5, 13 and 14 follow as being dependent on Claims 3 and 12.

Claims 1, 10, and the claims dependent thereon, are rejected as vague and indefinite because the "predetermined number" that provides a fixed length is unclear. From page 8, line 19 through page 9, line 2 of the specification, it is described that the databases of known biological fragments may be those that are published and publicly available, or that one can create a new motif database for comparison. On page 4, lines 20-21 of the specification, it is stated that the comparison database stores a predefined number of known biological sequences. On Specification page 9, lines 15-19, the terms "fixed number" or "total number" of motifs in the comparison database are used (to further explain or describe what is meant by "predetermined number"). Thus, it is evident that the "predetermined number" of biological fragments is the number of fragments or motifs in the published database, or in the database used as the comparison database. That number is a definite, well defined and accessible characteristic of the database. Further, base Claim 1 has been amended to recite this number as a "fixed number" of known biological fragments. Acceptance is respectfully requested.

Claims 1 and 10, and the claims dependent thereon, are rejected for not having a sufficient antecedent basis for the limitation "fixed length representation" of the genome sequence. Claims 1 and 10 have been amended to provide proper antecedent basis and to change "fixed length" to "uniform" as supported at least by Specification page 7, line 10.

Claims 6 and 15 are rejected as being unclear due to the contradiction between "subject genome sequence", which contains nucleic acids, and "protein sequence", which contains amino acids. These claims have been cancelled and replaced with new claims 19 and 20, which are written in independent form with changes to correct for the contradiction.

Claim 10, and the claims dependent thereon, are rejected as being vague and indefinite due to the methodology of output being unclear. Page 7, lines 20-23 of the specification clearly define the contemplated methods of output. The options include transmission of the output to a data file/data store, another program/processor routine, another computer coupled across a communication channel to a digital processor, and the like. Thus, the output contemplated by the present invention is clearly defined by the specification. Nonetheless, Claim 10 is now amended to recite that the "comparison routine provides the formed vector" (instead of "outputs the formed vector"). Dependent Claims 11-18 follow. Acceptance is respectfully requested.

According to the foregoing, all of the Examiner's rejections based on 35 U.S.C. § 112, second paragraph, have been addressed by either amendment or argument. No new matter is introduced. As such, Applicants believe that these claims satisfy the requirements of § 112 and it is respectfully requested that the § 112 rejections be withdrawn.

Rejections of Claims 1-4, 6-8, 10-13 and 15-18 under 35 U.S.C. § 102(b)

Claims 1-4, 6-8, 10-13 and 15-18 have been rejected under 35 U.S.C. § 102(b) as being anticipated by Levy *et al.*, "Xlandscape: the graphical display of word frequencies in sequences," *Bioninformatics*, Vol. 14 no. 1, pp. 74-80 (1998).

The "landscape" taught by Levy *et al.* is an array of variable size. Referring to Fig. 1 of the reference and the description of the algorithm on page 75, the landscape array has three rows and 15 columns. The array is lined up against a 15 base query sequence. The numbers in the first row represent the frequency of occurrence of the base below it. The numbers in the second row represent the frequency of a two-base sequence defined by the base below it and the next base. The numbers in the third row represent the frequency of a three-base sequence beginning with the base below it and the next two bases.

This type of landscape array differs substantially from the uniform vector representation claimed by Applicants. The Levy *et al.* landscape described in Fig. 1 is a comparison of a sequence of nucleotide bases with itself. It performs the function of graphically representing the frequency with which subsequences of varying length occur in the overall sequence. Thus, referring to the third row, third column of Fig. 1 of Levy *et al.*, there is a 2, which means that the three-base sequence "tcc" appears two times in the whole sequence. Levy *et al.* refer to this as

the special case where a query sequence equals a database and suggest that a database other than the query sequence can be used to indicate matches between the query sequence and a database.

Page 9 of the Office Action at hand presents a table which represents the values of each cell of the array arranged in a table that is m columns by n rows and argues that each m and n are a predetermined length, and thus a fixed length. Applicant respectfully disagrees for reasons set forth below.

Base Claims 1 and 10, as amended herein, recite that the resulting vector is "a uniform representation of the subject genome sequence." The reason that it is uniform is because its length is determined by the fixed number of known biological fragments in the provided set. Thus, when a second subject genome sequence is compared to the same provided set of known biological fragments, the resulting vector will have the same length as the vector of the first subject genome sequence, as will the vectors corresponding to all subsequent subject genome sequences, hence a uniform vector representation. The resulting vector, if expressed as an array to be analogous to the Examiner's interpretation of Levy *et al.*, would have as many rows as there were biological fragments in the provided set. For any given row, the value would be the frequency with which a particular biological fragment occurs in the subject genome sequence. The size of the Levy *et al.* array, in contrast, depends entirely on the size of the subsequences which appear frequently in the sequence.

To illustrate, consider the sequence shown in Fig. 1 of Levy *et al.* It's resulting array, or landscape, is 3 rows by 15 columns. The reason that it is 3 rows is because the largest subsequence that appears repeated in the sequence is 3 bases ("tcc"). Consider a second sequence in which a 4-base subsequence was repeated. In this case, the array would have been 4 rows in depth. In fact, with 15 bases, the array could be as deep as 7 rows. That this is a sequence compared to itself is irrelevant. If it were the representation of a sequence compared to a database or other sequence, the variability of the number of rows would still depend on the length of the largest subsequence which appears in the sequence. In that case, the array could be as much as 15 rows deep. Thus, Levy *et al.* is not teaching any form of uniform length representation of the subject sequence.

Applicant's vector representation, in contrast, will always have the same number of rows for any subject genome sequence which is compared to the same set of known biological fragments because the number of biological fragments in that set does not change.

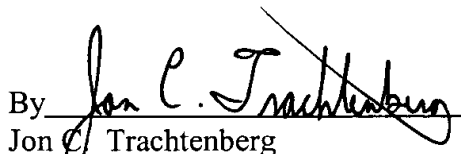
Thus, Levy *et al.* fails to teach or suggest a vector of "uniform representation of a subject genome sequence" and as such, cannot anticipate Applicants' invention as now claimed in base Claims 1 and 10 and new Claims 19 and 20. Dependent claims 2-4, 7-8, 9-13 and 16-18 (claims 6 and 15 being cancelled) follow. Accordingly, the § 102 rejection Levy *et al.* is believed to be overcome.

CONCLUSION

In view of the above amendments and remarks, it is believed that all pending claims (Claims 1-5, 7-15, 16-20) are in condition for allowance, and it is respectfully requested that the application be passed to issue. If the Examiner feels that a telephone conference would expedite prosecution of this case, the Examiner is invited to call the undersigned at (978) 341-0036.

Respectfully submitted,

HAMILTON, BROOK, SMITH & REYNOLDS, P.C.

By 

Jon C. Trachtenberg
Registration No. 45,455
Telephone: (978) 341-0036
Facsimile: (978) 341-0136

Concord, MA 01742-9133
Dated: 12/23/02

MARKED UP VERSION OF AMENDMENTS

Specification Amendments Under 37 C.F.R. § 1.121(b)(1)(iii)

Replace the paragraph at page 8, line 19 through page 9, line 2 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

The BLOCKS database (Steven Henikoff and Jorja G. Henikoff, "Automated assembly of protein blocks for database searching," *Nucleic Acids Research, 19:23*, pp. 6565-6572 (1991)) is an example of a database 17 of motifs. Emitof [(http://dna.stanford.edu/emotif/)], and PRINTs [(http://bioinf.man.ac.uk.dbbrowser/ PRINTS/)] are other such databases. These and other published databases may be used as the working predefined set/comparison database 17 in the present invention. Alternatively, it is possible to create a new motif database 17 from any protein database which has been labeled according to some parameter (e.g., structure). This is achieved by using multiple alignment software to find short multiply aligned ungapped sequences and then collecting statistics about these in a matrix [(http://www2.ebi.ac.uk/clustalw/, http://www.blocks.fhcrc.org/)]. By creating a motif database 17 specific to the proteins of interest 11, more meaningful feature vectors 23 may be obtained since the motifs from a more general database may not occur in the proteins of interest.

Claim Amendments Under 37 C.F.R. § 1.121(c)(1)(ii)

1. (Amended) A method for [uniform representation of] analyzing a subject genome sequence comprising the steps of:

providing a set of known biological fragments, the set being of a [predetermined] fixed number of said known biological fragments, each known biological fragment in the set having a respective representation;

comparing the respective representation of each known biological fragment from the set to a subject genome sequence, for each known biological fragment said comparing including (i) counting the number of times the respective representation of the known biological fragment is found in the subject genome sequence and (ii) from said counted number of times, forming a vector element, such that for each known biological fragment there is a respective vector element representing the number of times the respective representation of that known biological fragment is found in the subject genome sequence; [and]

from the formed vector elements, forming a vector having a length equal to the [predetermined] fixed number of known biological fragments in the provided set, such that the formed vector provides a [fixed length] uniform representation of the subject genome sequence; and

providing the formed vector for use as input to a desired analysis, the uniform representation provided by the formed vector enabling the formed vector to serve as normalized input.

3. (Amended) A method as claimed in Claim 1 further comprising the step of:

for each desired subject genome sequence, using said set of known biological fragments, repeating the comparing and forming steps such that a respective vector representation is formed and each desired subject genome sequence has a [same length] respective vector representation of a same length, said set of known biological fragments being a same set used for all of said subject genome sequences.

10. (Amended) Apparatus for [forming uniform representations of] analyzing genome sequences, comprising:

a data store of a predefined number of known biological sequences;

a comparison routine executed by a digital processor having access to the data store, the comparison routine comparing each known biological sequence from the data store to a subject genome sequence and generating a score indicative of the comparison, said scores forming a vector having a length equal to the predefined number of known biological

sequences, such that said comparison routine [outputs] provides the formed vector as a [fixed length] uniform representation of the subject genome sequence and the formed vector enables at least one analysis of the subject genome sequence, the uniform representation of the formed vector providing normalized input for the analysis.

12. (Amended) Apparatus as claimed in Claim 10 further comprising a plurality of different subject genome sequences; and

wherein, [the comparison routine] using a same set of known biological sequences, the comparison routine forms, for each subject genome sequence, a respective vector such that a corresponding plurality of [same] uniform length vector representations is provided.

13. (Amended) Apparatus as claimed in Claim 12 wherein the output of the comparison routine feeds the corresponding plurality of [same] uniform length vector representations into further analysis processors.